Towards a Thesaurus for Energy Efficiency in Building Construction: the Italian context*

Elena Cardillo[1], Antonietta Folino[2], Francesca Iozzi[2],
MariaTaverniti[2] and Elisabetta Oliveri [3]

[1] Fondazione Bruno Kessler, FBK-ISRT, via Sommarive 18, 38100,
Povo, Trento (Italy), fax: 0461 302040
cardillo@fbk.eu
[2] Laboratorio di Documentazione, Universita' della Calabria, via P.Bucci cubo 20B, 87036,
Arcavacata di Rende, Cosenza (Italy), fax: 0984 494625
{antonietta.folino, francesca.iozzi, maria.taverniti}@unical.it
[3] ITC-CNR Istituto per le Tecnologie della Costruzione, Viale Lombardia, 49, 20098,
San Giuliano Milanese, Milano (Italy), fax: 02 98280088
elisabetta.oliveri@itc.cnr.it

Abstract: During the last few years, improvements in Renewable Sources to Building Construction, contributed to intensify the terminological problem, where are evident a high level of heterogeneity, and the lack of a standardized and unambiguous vocabulary/classification system. This is due to many factors, such as building techniques and working tools evolution, technical law evolution in Italy and in Europe. In this paper we want to focus on the Italian context, proposing a methodology to face this problem, which is based on the creation of a specialist lexicon and a Faceted Thesaurus for the specific domain. These will allow to organize and structure the domain of sustainable construction and energetic saving, and to improve access to up-to-date and unambiguous information in a web-based environment.

Keywords: Specialist Lexicon, Thesauri, Faceted Classification, Building Construction, Renewable Energies.

1 Introduction

The field of Renewable Energies applied to building construction involves lots of stakeholders. The fact that in our country there are no international «gold standard» systems for classifying and sharing such information causes problems above all related to the missing data~~.~~

To face with this problem we want to create a Knowledge Organization System (KOS), in particular a Faceted Thesaurus, for all the users involved in the different domains of the Italian construction sector.

*This work is supported by the Project *«Nuove tecnologie e strumenti per l'efficienza energetica e l'utilizzo delle fonti rinnovabili negli usi finali civili»*, promoted by the Regione Lombardia and by the Italian National Council of Research.

In fact, there is no uniform system for classifying information/terminology relating to construction products, materials, services and machineries and there is no standardized classification of construction cost[1]. This inhomogeneous situation causes the production and adoption of several not standardised systems of classification. Moreover, the necessity of creating a KOS for such a domain grows from the need of facilitating access to and management of specialist knowledge to experts of the field, but also to less specialized users.

The choice of adopting the faceted technique, for constructing a thesaurus in our specific domain, consists in the possibility of using it as a user interface for searching, which would allow access to a knowledge repository filled in with the domain knowledge and a specialist vocabulary. In fact, such resources will be integrated in a web portal based on an open source documental management tool, and they will allow the access to the information organised in such a system.

2 Background

To overcome the situation that characterizes our country, a UNI project named *Edilizia e Opere di Ingegneria Civile. Criteri di codificazione di opere, attività e risorse - Identificazione, descrizione e interoperabilità*, has been proposed, in March 2009, by the *Regione Lombardia* jointly with other important associations in the construction domain. This project aims to create a system of unique codes in order to have a univocal way of description for identify each material or service facilitating the sharing of the information among all the stakeholders and in order to establish a dialogue between databases containing information.

For the Building Construction domain a non exhaustive sample of controlled vocabularies, could be the following: the ICONDA Terminology (Fraunhofer IRB)[2] , ISO12006 parts 2 and 3[3], the LexiCon (the Netherlands), Barbi (Norway), BC Building Definitions taxonomy (e-Construct IST Project), BS6100 and UNICLASS (British Standards)[4], e-COGNOS ontology (e-COGNOS IST project), and the Standard Dictionary for Construction in France (SDC).

In other continents similar efforts were also conducted, such as the CI/SfB[5], Masterformat, Omniclass, and the Canadian Thesaurus, just to name a few. Other two system of classification related with Uniclass are: CAWS (Common Arrangement of Work Sections for building works) and EPIC (Electronic Product Information

[1]http://prezzari.str.it/Pagine/ElencoListini.aspx.

[2]http://library.dialog.com/bluesheets/html/bl0118.html.

[3] ISO 12006-2 (2001) Organisation of Information about Construction Works- part 2: Framework for Classification of Information; cfr. Anders Ekholm, *ISO 12006-2 AND IFC – Prerequisites for coordination of standards for Classification and Interoperability,* ITcon Vol. 10, p. 275-289, http://www.itcon.org/2005/19.

[4] http://www.cpic.org.uk/en/publications/uniclass-listing.cfm.

[5] CI/SfB stands for Construction Index/Samarbetskommitten for Byggnadsfragor. This is a faceted classification system specially designed for the construction sector.

http://www.ascinfo.co.uk/9/category/category13_9.html.

Cooperation), both are systems for structuring product data and product literature. However, more in detail, OmniClass Construction Classification System[6] (the American's equivalent of Uniclass) is a faceted classification system for the Construction Industry, based on ISO 120062 and ISO/PAS 12006-3[7] and follows the international framework set out in ISO Technical Report 14177 *Classification of Information in the Construction Industry July 1994.* OmniClass^{tm} incorporates in it: MasterFormat^{tm} for work results, UniFormat^{tm} for elements and EPIC for products. These and other similar initiatives can be found in (Lima, *et al.*, 2007). Another important project is SEAMLESS[8], developed in 2006. The project aims to construct a global ontology of the Building and Construction sector (B&C GLOB).

In conclusion, controlled vocabularies construction (Aitchinson *et al*., 2000; Broughton V., 2008) is regulated by the BS 5723 (1987), ISO 2788 (1986), ANSI/NISO Z.39 2005 – for monolingual thesaurus - and the ISO 5964 (1985) – for multilingual thesaurus.

3 Motivation and Approach

3.1 The problem of Terminology in the field of Building Construction

In our country, the fields of energy efficiency and of the application of renewable sources to the construction of buildings are characterised, from a terminological point of view, by a high level of heterogeneity. Several factors contribute to this situation: influences coming from local languages, and local building traditions and materials; language, building techniques, working tools and technical law evolution; adoption of linguistic loans.

A lack of homogeneity can be also found in the way entities are classified: a unique classification system does not exist in our national context, so the same object could be found under different categories, with obvious problems in the characterization and in the search of the same objects. One of the possible causes could be found in the erroneous use of certain terms that, as a consequence, are classified under the wrong category. A significant example is the classification used by each region in their own prices catalogues: it should be important to have a certain level of coherence in such documents, because they should provide a common reference for a budget estimate in the realization of building works. Differences in this and in other kinds of documents could cause economic and security problems.

[6] http://www.omniclass.org/.

[7] ISO 12006-2 (2001) Organisation of Information about Construction Works- part 2: Framework for Classification of Information; cfr. Anders EKHOLM, *ISO 12006-2 AND IFC – Prerequisites for coordination of standards for Classification and Interoperability,* ITcon Vol. 10, pg. 275-289, http://www.itcon.org/2005/19; ISO/PAS 12006-3 (2007), *Organisation of Information about Construction Works- part 3: Framework for Object-Oriented Information*.

[8] http://www.seamless-eu.org/home.html.

3.2 Approach and Contributions

The methodology chosen to reach the mentioned objectives is based on the creation of an Italian thesaurus for the domain of construction technologies. This activity includes several steps.

3.2.1 Corpus Construction and Markup Process

The corpus on which the process of terminology extraction will be performed to find representative and unambiguous terms of the domain is constituted by hundreds of documents such as: laws and technical standards, scientific reviews, books, guidelines, grey literature, technical documents from production companies. All documents are inserted in the document management platform after having been structured and described by the elements of the standard Dublin Core (DC). Documents markup is based on the markup language XMP (eXtensible Metadata Platform)[9], which is based on DC metadata schema and on the XML standard. These two principal characteristics have determined the choice of this formalism: 1) metadata are embedded in the file, 2) XMP metadata can describe a document as a whole, but can also describe parts of a document, such as pages or included images.

3.2.2 Term Extraction and Statistical Analysis

A semi automatic term extraction will be performed on the created corpus, using the dedicated software T2K (Text-2-Knowledge), a tool developed at the ILC (Institute of Computational Linguistic) of Pisa[10]. This program performs a linguistic analysis on the texts and then it provides, as final output, a term-based vocabulary, where terms are organized in a hierarchical hyponym/hyperonym relation depending on their internal linguistic structure. The candidate terms detected by T2K can be either single or multi-word terms, and represent the terminology index of the domain analysed. The extracted terms related to the domain of interest will be manually included in the thesaurus. This process will be followed both for its construction and for its updating. Even if this process seems to be difficult, particularly in case of updating, it appears to be a coherent solution, at least in the current state of our work. Automatic extraction will allow detecting new terms related to the domain and the complexity of the thesaurus structure requires a manual registration. Nevertheless, the possibilities of automation in information extraction and in dynamically categorization of terms by the use of metadata require further evaluation.

3.2.3 Creation of the Specialist Lexicon for the domain

On the basis of the Term Extraction process a Specialist Lexicon for the domain will be realized. It should be shared by all the subjects that participate to the creation and use of information. This resource will be strictly related to the Thesaurus we want

[9] Formalism created by the Adobe Society in 2001: http://www.adobe.com/products/xmp/overview.html.

[10] http://www.ilc.cnr.it/indexnoflash.html.

to create, because the vocabulary will contain those terms that in the thesaurus will be treated as descriptors or preferred terms.

3.2.4 Faceted Thesaurus Construction

Our choice to create a Thesaurus resides in the fact that it offers the possibility to structure information, to control the variability of language and in part to solve terminological incongruities by relations between terms, in particular, by the relation between preferred and non-preferred terms.

The regional and local variants of descriptors and the more commonly used terms will be introduced in the thesaurus as non preferred terms or cross-references, and they will be related to descriptors as their variants or synonyms by the equivalence relation. This organization should provide an easier access to information to different types of users: professionals in the field of construction; fitters and maintainers; production companies; universities and research institutes; students; public organizations. This thesaurus should be designed on the base of the users' demand: they belong to different competences, so terms used to search for information will not be the same for each one of them. The relationship network between terms in the thesaurus should allow users obtaining relevant information even if the term used for the research doesn't correspond to the preferred term. Relations between variants, synonyms, etc. should be clearly showed to users, as well as differences between similar terms.

The role of the user will consist in navigating and in formulating queries. At the moment we haven't envisaged the possibility that users could actively contribute to the thesaurus construction, by adding new terms. So the choice and the organization of attributes will be established and improved by professionals in information management, supported by experts of the domain.

Thanks to its systematic structure, a thesaurus represents a sort of classificatory structure, where terms can be organized according to a series of categories.

The classification system we want to propose would not be based on a particular scheme among those that already exist, but should be constructed from scratch, on the basis of representative documents, in particular laws and technical rules. A simple transposition from the existing international classifications is not possible: each country has its own classification criteria, depending on factors such as climate, building techniques and similar ones that cannot be used in another country and that influence terminological and classification choices.

From a methodological point of view, it would be appropriate to use a flexible and multidimensional scheme, like the *faceted* one, in order to classify the various objects, terms and documents, on the basis of several criteria, representing the properties of the objects themselves. The advantages provided by a faceted thesaurus consist in a more efficient research activity and in an easier possibility of updating.

The identification of the classification schema could be based on the general standard facets proposed by the Classification Research Group (CRG), appropriately adapted to the specific context. Some of these categories could be for example *technologies*, differentiated in «passive», like *thermal insulation* and «active», like *solar thermal*, *materials*; *reinforced concrete*, *products*; *solar panels*, *organizations*, and other similar categories.

4 Concluding Remarks

In this article we presented a proposal about a possible methodology of constructing a thesaurus in the domains of Energy Efficiency and of the application of the renewable sources to the construction of buildings. We have proposed an approach that could represent a potential solution for such problems. Being still in progress, the presented project hasn't produced evaluable results, yet.

Nevertheless, the expected ones consist in the opportunity to provide our country with a unique way of organize information about the specific field, with an easier and more pertinent way of accessing up-to-date information and by an unambiguous collection of specialist terms.

References

AITCHINSON J., GILCHRIST A., BAWDEN D., (2000). *Thesaurus construction and use: a practical manual*, Europe Publications, London.

BRITISH STANDARD INSTITUTIONS (BS). (1987). BS 5723:1987 *Guide to establishment and development of monolingual thesauri, UK. The British Standard Institutions.*

BARTOLINI R., LENCI A., MARCHI S., MONTEMAGNI S., PIRRELLI V., (2005). *Text-2-knowledge: Acquisizione semi-automatica di ontologie per l'indicizzazione semantica di documenti*, Relazione al progetto PEKITA, ILC. Pisa p.23.

BROUGHTON V., (2008). *Costruire Thesauri*, edited by Piero Cavaleri. Translated by Laura Ballestra e Luisa Venuti, EditriceBibliografica.

EKHOLM A., (2005), *ISO 12006-2 AND IFC – Prerequisites for coordination of standards for Classification and Interoperability,* ITcon Vol. 10, pg. 275-289, http://www.itcon.org/2005/19.

GOH B. H, CHU Y. L., (2002), D*eveloping National Standards for the Classification of Construction Information in Singapore.* In Proceedings of the International Council for Research and Innovation in Building and Construction conference.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO). (1986). ISO 2788:1986 Documentation -- Guidelines for the establishment for the construction of monolingual thesauri. Geneva, Switzerland: International Organization for Standardization.

LIMA C., ZARLI, A., STORER, G., (2007). *Controlled Vocabularies in the European Construction Sector: Evolution, Current Developments, and Future Trends,* in Complex Systems Concurrent Engineering, Springer London.

NATIONAL INFORMATION STANDARDS ORGANIZATION (NISO). (2005). *Guidelines for the construction, format, and management of monolingual controlled vocabularies*: ANSI/NISO Z39.19-2005, Bethesda Md., NISO Press.